August 2021
Master's Degree Thesis

Reinforcement Learning-based User Association for Cloud Radio Access Networks

Graduate School of Chosun University

Department of Computer Engineering

Rehenuma Tasnim Rodoshi

Reinforcement Learning-based User Association for Cloud Radio Access Networks

클라우드 무선 접속 네트워크를 위한 강화학습 기반 사용자 접속 기술 연구

August 27, 2021

Graduate School of Chosun University

Department of Computer Engineering

Rehenuma Tasnim Rodoshi

Reinforcement Learning-based User Association for Cloud Radio Access Networks

Advisor: Prof. Wooyeol Choi

A thesis submitted in partial fulfillment of the requirements for a Master's degree

April 2021

Graduate School of Chosun University

Department of Computer Engineering

Rehenuma Tasnim Rodoshi

로도시 레헤누마 타스님
석사학위논문을 인준함

위원장 조선대학교 교수 신석주 (인)

위 원 조선대학교 교수 강문수 (인)

위 원 조선대학교 교수 최우열 (인)

2021년 05월

조선대학교 대학원

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

C-RAN	Cloud Radio Access Networks
BBU	Base Band Unit
RRH	Remote Radio Head
mmWave	Millimeter wave
QoS	Quality of Service
RL	Reinforcement Learning
FL	Fuzzy logic
TTT	Time-to-trigger
SNR	Signal-to-noise Ratio

List of Figures

List of Tables

ABSTRACT

Reinforcement Learning-based User Association for Cloud Radio Access Networks

Rehenuma Tasnim Rodoshi
Advisor: Prof. Wooyeol Choi, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

Cloud radio access network (C-RAN) is a promising architecture for the 5G mobile communication system that provides seamless connectivity to the users while satisfying the ever-increasing user demand. In C-RAN, the base station functionality is divided into baseband unit (BBU) and remote radio head (RRH), then the BBUs from multiple sites are centralized and virtualized using cloud computing and virtualization techniques. All the data processing and controlling are performed inside the BBU pool and RRHs are responsible for radio functionalities. According to the requirements of 5G, the short-range small cell-based RRHs are densely deployed in an overlapping manner. The mobility of users has a significant impact on their association with RRHs when a user moves within the coverage of multiple RRHs. The traditional handover schemes mostly rely on the signal strengths a user receives from an RRH which will cause a large number of unnecessary and frequent handovers. So, it is necessary to optimize handover control parameters before the handover occurs for a user and re-associate the user to an RRH that reduces the unnecessary handovers in the

network. This paper investigates the handover in C-RAN by carefully optimizing the handover control parameter with fuzzy logic and selecting the target RRH for handover with a reinforcement learning (RL) algorithm. A key ingredient of the proposed RL-based scheme is to use an acceleration technique for faster convergence of the algorithm. Our main goal is to re-associate users with an RRH in a way such that the association after the handover remains as long as possible while maintaining the quality of service (QoS) requirements of the users. Numerical results show that the proposed scheme can significantly reduce the number of handovers while ensuring the QoS requirements.

한 글 요 약

클라우드 무선 접속 네트워크를 위한 강화학습 기반 사용자 접속 기술 연구

레헤누마 타스님 로도시
지도 교수: 최우열
컴퓨터공학과
대학원, 조선대학교

클라우드 무선 액세스 네트워크(C-RAN)는 끊임없이 증가하는 사용자 수요를 충족시키면서 사용자에게 원활한 연결을 제공하는 5G 이동통신 시스템 아키텍처이다. C-RAN에서 기지국 기능은 BBU(Base Band Unit)와 RRH(Remote Radio Head)로 구분되며, 그 다음 여러 사이트의 BBU는 클라우드 컴퓨팅 및 가상화 기술을 사용하여 중앙 집중화되고 가상화된다. 모든 데이터 처리 및 제어는 BBU pool 내에서 수행되며 RRH는 무선 송수신 기능을 담당한다. 5G의 요구 사항에 따라, 단거리 소형 셀 기반 RRH가 중복 배치된다. 사용자의 이동성은 사용자가 여러 개의 RRH 범위 내에서 이동할 때 RRH와의 연결에 상당한 영향을 미친다. 기존의 핸드오버 방식은 대부분 사용자가 RRH로부터 수신하는 신호 강도에 의존하여, 불필요하고 빈번한 핸드오버를 유발한다. 따라서 사용자가 핸드오버를 수행하기 전에 핸드오버 제어 매개 변수를 최적화하고, 사용자를 네트워크에서 불필요한 핸드오버를 줄이는 RRH에 다시 연결해야 한다. 본 논문은 fuzzy logic을 이용하여 핸드오버 제어 매개변수를 신중하게 최적화하고, 강화학습 알고리즘을 사용하여 핸드오버를 위한 대상 RRH를 선택하여 C-RAN에서의 핸드오버를 수행한다. 제안된 강화학습 기반 사용자 선택 방법의 핵심 요소는 알고리즘의 빠른 수렴을 위해 가속 기술을 활용한다. 본 연구의 주요 목표는 사용자의 서비스 품질(QoS) 요구사항을 유지하면서 핸드오버 후

연결성이 최대한 오래 유지되도록 사용자를 RRH와 다시 연결하는 것입니다. 다양한 환경에서의 시뮬레이션 결과는 제안된 방법이 QoS 요구사항을 보장하면서 핸드오버 횟수를 크게 줄일 수 있음을 보여준다.

I. INTRODUCTION

A. Overview

Due to the growing number of mobile and internet of things (IoT) devices, one of the critical challenges facing by mobile communication network is that of satisfying the increasing traffic demands such as high data rate and better quality of service (QoS) with low latency [1]. Fifth-generation (5G) mobile network is envisioned to provide a high data rate with massive connectivity and mobility support exploiting the millimeter wave (mmWave) band for communication. However, mmWave communication suffers from significant sensitivity to blockage and a high penetration loss due to the short wavelength [2]. In order to overcome these challenges in mmWave, small cell technology is introduced in 5G which has a shorter coverage range. MmWave communication can be provided within the range of small cells so as to avoid obstacles and reduce signals to get easily blocked. 5G supports small cells of different coverage ranges (microcell, picocell, femtocell) deployed in an overlapping manner that co-exists with the existing LTE macro cell [3]. By increasing the number of small cells, a higher data rate, and massive connectivity for the exponentially rising number of devices can be provided.

One of the promising mobile network architectures that match the core features of 5G for enhancing network capacity with seamless connectivity is the cloud radio access network (C-RAN) [4]. The architecture of C-RAN for 5G communication scenario is given in Figure 1. In C-RAN, the base station is divided into baseband unit (BBU) and remote radio head (RRH), and then the BBUs from multiple sites are centralized and virtualized using cloud computing and virtualization techniques [5]. The centralized and virtualized architecture of

C-RAN gives the advantages of adapting to dynamic traffic fluctuation achieving load balancing, cost reduction, and interference minimization. In C-RAN, RRHs are connected to the BBU pool through the fronthaul link. The inter-RRH interference is mitigated by joint coordination through the centralized cooperative processing in the BBU pool [6]. However, the number of users one RRH can support at a certain time is restricted due to the limited fronthaul capacity.

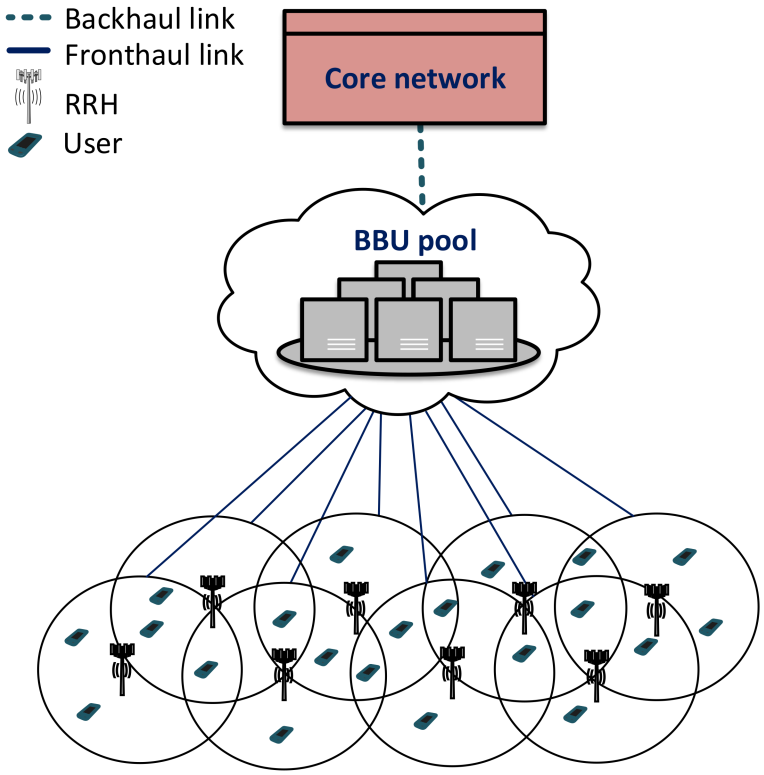

Figure 1: The architecture of C-RAN with virtualized BBU pool and small cell-based RRH for 5G communication

B. Research objective

As expected in 5G, the RRHs of C-RAN are densely deployed in an overlapping manner, and users are moving at different speeds. While the dense deployment of RRHs enhances the data rate and connectivity, it will lead to more frequent switching of user association from one RRH to another. This procedure of switching user association between RRHs or re-association of users is known as handover. User mobility has a significant impact on the number of handovers because it may result in frequent handovers as users move from the coverage of one RHH to another within a short time. In a certain location, a user can be in the coverage of more than one RRH and it may receive a high signal from multiple RRHs. Traditional cellular network handover policy is based on the received signal strength which is not suitable for small cell-based C-RAN in 5G. In that case, the user association may be changed frequently causing unnecessary handovers in the network. Frequent handovers lead to heavy signaling overhead, low energy efficiency, and reduced throughput in the network. Along with the received signal, many other control parameters need to be considered to develop an efficient handover mechanism. It is necessary to execute the handover effectively so that the connection with the serving RRH does not change frequently as well as during handover user gets connected to an RRH for which the connection remains longer.

Various techniques have been proposed in the literature for effective handover management and successful re-association of users to RRH. In this regard, different parameters have been used to reduce the number of handovers in the network. According to 3GPP [7], six handover events and two handover control parameters have been defined. Considering different events, the handover control

parameters are adjusted to control the handover trigger condition. Along with the received signal strength, considering distance, velocity, direction has shown better performance for handover management. Furthermore, the association of users with a suitable base station (RRH) has been optimized by considering various parameters based on the objective of the studies.

Although handover control parameter optimization and suitable RRH selection have been studied in the literature, it is necessary to integrate both optimizations in a single study to maintain network efficiency. Moreover, the parameter selection for optimizing the handover trigger condition is very crucial. The parameters must reflect the objective of reducing the number of handovers while maintaining the connection with the minimum data rate. Due to the mobility of users, a user may lose the connection at the next time-stamp even if the received signal is strong at the current time. So, the parameters have to be chosen in a way so that the possible location at the next time can be known. Furthermore, the candidate RRHs selection for association with a user during handover has to be performed properly. The target RRH to which the user will be connected is selected from the set of candidate RRHs. Rather than performing a short-sighted RRH selection for user association, it has to be done in a way so that the selected RRH maintains the connection for a longer duration in order to reduce the overall number of handovers.

C. Contributions

In this study, we investigate the user re-association problem in C-RAN and provide a solution to minimize the number of frequent handovers while maintaining the QoS requirement of users. Our proposed scheme consists of two

parts. First, we determine the handover trigger condition and optimize a handover control parameter called time-to-trigger (TTT). TTT is the duration for which the connection between a user and RRH remains after the received signal strength becomes less than a threshold value. In order to initiate a handover, along with the received signal strength of the user, we have considered the velocity of user and the distance between user and RRH. The handover triggering condition is optimized using these parameters with the fuzzy logic process. Based on the decision of fuzzy logic, when the condition is satisfied, an RRH is chosen for the user for association. We have utilized an RL-based algorithm to select the RRH in a way such that the connection will remain longer. In order to make the RL algorithm converge faster, we have proposed a prediction-based virtual reward creation and mapping of virtual reward with the actual reward. The novelty of this work lies in optimizing both the handover triggering and RRH selection during handover for re-association. Moreover, in order to accelerate the learning of the RL algorithm for user association, we utilize a prediction-based virtual reward updating. This acceleration technique helps in faster convergence with performance improvement.

The contribution of this study can be summarized as follows:

- The user re-association with RRH is investigated with an objective to minimize the number of frequent and unnecessary handovers in the network while satisfying the QoS of users. In this regard, the handover triggering and RRH selection for user association have been optimized.

- The handover triggering condition is optimized by adjusting the time-to-trigger (TTT) values using fuzzy logic considering the received signal, distance, and velocity of the user. This ensures that no early handover occurs in the

network while maintaining the connection. Also, the candidate RRH selection is performed in a way to reduce the ping-pong handovers.

- After the handover trigger condition is met, an RL algorithm is used to choose the target RRH for users, aiming to keep the user-RRH association for as long as possible. The state space is constructed based on the user and RRH information, the action is to select the RRH for association and the reward function reflects the objective of our work.

- In order to accelerate the convergence of the RL algorithm, we have proposed a prediction-based virtual reward creation along with the actual reward in certain conditions. The exploration-exploitation strategy of RL is designed to take advantage of the acceleration technique.

- To evaluate the performance of the proposed scheme, a simulation is done for the C-RAN environment. We verify both the fuzzy logic-based handover parameter optimization and RL-based RRH selection with acceleration technique (FLRL_ac). According to our evaluation results, the proposed FLRL_ac technique outperforms the conventional scheme in terms of the number of handovers per user and average connection remaining time for the user-RRH association

D. Thesis layout

The rest of the thesis is organized as follows:

In Chapter 2, the related works are reviewed and discussed along with the limitation of the existing studies. Chapter 3 provides the system model of the proposed scheme including the assumptions made in this study. In Chapter

4, the handover framework is given at first describing the optimization of handover trigger condition with fuzzy logic. Then, the user association scheme with RL-based RRH selection is described. The state, action, and reward of the RL algorithm are defined and the prediction-based virtual reward creation procedure is also discussed. In Chapter 5, the performance of the proposed method is evaluated and compared with other user association schemes. Finally, the conclusion is given in Chapter 6.

II. RELATED WORKS

While handover management and user association have been investigated widely in literature, very few studies have optimized both handover control parameter and user association simultaneously. Moreover, none of the related studies considered solely the C-RAN architecture, although centralized controller or software-defined networking (SDN)-based handover or user association has been investigated in a few of the works. This section first discusses the related studies on handover parameter optimization and user association schemes separately. Then, the research gap in the literature and our main contributions for filling up the gap are highlighted.

A. Handover parameter optimization

Due to the dense deployment of small cells in 5G, handover control parameter optimization has become of great concern. Many researchers proposed different methods for adjusting handover parameters to initiate handover with different objectives such as minimizing handover failure ratio, handover delay, average number of handovers, ping-pong handovers (user is handed over back and forth from the serving RRH and the target RRH over a short period), and frequent handovers. Most of the research works related to handover parameter optimization are proposed based on LTE communication [8], [9]. These methods are not suitable for 5G communication due to the integration of small cells in 5G. For the literature review, we discuss the handover parameter optimization schemes based on 5G or centralized network architecture, which are more related to the objective of our study.

In [10]–[12], handover management has been discussed and different types of optimization techniques have been proposed with different objectives. The

parameters considered for handover management are also different based on the objectives. A weighted fuzzy self-optimization scheme has been proposed in [10] to optimize two handover control parameters: handover margin and TTT using the value of SINR, the traffic load of serving, and target base station, and user velocity. Their objective was to lower the rates of ping pong handovers, and radio link failure.

Enhancing the work in [10], the authors in [11] proposed a distributed velocity-aware algorithm to optimize the two handover control parameters handover margin and TTT. The value of RSRP and user velocity is measured with a threshold value to make the handover decision. The main performance metrics considered were handover probability and ping pong handovers. Although these studies optimized two handover control parameters, the target base station is optimized only based on the traffic load. Other parameters such as distance and direction between user-RRH are necessary to select the target base station so as to determine the possible connection duration. A target base station with a low traffic load can provide a high data rate at that time, but it does not provide any view of how long the user will be connected. Moreover, the optimization does not consider reducing frequent handovers or unnecessary handovers.

The authors in [12] proposed a fuzzy logic-based method to minimize ping pong handovers and handover failure ratio in dense small cell networks. They took into consideration user velocity, RSRP, and RSRQ to get a value of RSRP as output for initiating the handover. However, the target base station selection is not optimized in this work. Also, the proposed method does consider the amount of time the user may stay under the serving or a target RRH.

B. User association for reducing handover

Several studies have been performed for associating a user to a base station or RRH with an objective to reduce the number of handovers in the network. We discuss the base station selection or user association schemes to reduce handover proposed in the literature.

In [13], a user association strategy in the mmWave ultra-dense network has been proposed to select the optimal base station that maximizes the user-BS association duration. An offline double deep reinforcement learning has been utilized to make the handover decision by mapping SNR values to the UE trajectory. The main objective was to reduce the number of handovers considering HO cost and increase the system throughput to mitigate the adverse QoS. Considering only the SNR values for selecting the target base station does not ensure the longest connection duration.

Some studies have been done to optimize handover with the help of a centralized controller [14], [15]. The controller observes the user and necessary measurements, thus select an optimal base station or time of handover occurrence, similar to C-RAN. In [14], an optimal user association scheme for uplink data transmission in C-RAN is proposed to reduce handover while balancing load between base stations. The association solution is reusable, which means the same association can be repeated if no changes occur in the network topology. Although this study reduces the number of handovers while performing load balancing, this does not ensure selecting the RRH with the longest connection duration.

Fang et al. utilized a dynamic particle swarm optimization-based algorithm in the central controller for handover management in [15]. Their main objective

was to optimize the overall quality of experience (QoE) of users and minimize the proportion of users with extremely low QoE. Bilen et al. in [15] investigated the handover management for SDN-based ultra-dense 5G network to reduce handover delay. The mobile node mobility and available resource in base stations are estimated using the Markov chain. Using the estimated values, the optimal base stations are chosen and assigned to the mobile nodes virtually prior to the need for an actual connection.

RL-based user association in 5G communication for reducing handover has been proposed in [16], [17], which is more related to our proposed mechanism. However, none of the works consider the specific properties of C-RAN architecture. In [16], the handover trigger condition is determined by considering the data rate of users using a threshold value. No other parameters such as distance, velocity are used for optimizing handover trigger condition. Moreover, the design of the RL algorithm is different in SMART than in our work.

Similarly, the RL-based base station selection algorithm in [17] does not optimize the handover trigger condition, although the optimization objective is the same for target base station selection. An acceleration technique for faster convergence of RL has also been proposed in this work. But the virtual reward may not be accurate as the exact association duration cannot be known. So, a prediction of the duration and providing a virtual reward based on the prediction can be effective.

Unlike the above-mentioned techniques, we aim to reduce the number of handovers while maintaining the QoS of users by optimizing both handover trigger condition and target RRH selection for the user. The main goal is to associate the user with an RRH in which the user will stay for a long duration

while getting the minimum required data rate. None of the above-mentioned works performed handover parameter optimization and target RRH selection simultaneously. Also, the RL-based RRH selection algorithms are not proposed for specifically C-RAN architecture, in which the BBU pool carries out the user association algorithm. Moreover, along with the total number of handovers per user as evaluation metrics in our study, the average user-RRH association duration is also considered, which more directly reflects the quality of the handover decision.

III. SYSTEM MODEL

In this chapter, we discuss the assumptions, initial user association, propagation model, and QoS model. The symbols and notations used in this study are given in Table 1

We consider a C-RAN architecture consisting of m mmWave small RRHs densely deployed in the network. RRHs are distributed in an overlapping manner that increases the overall network capacity while minimizing out-of-service areas. The set of RRHs can be denoted by M where $M = \{1,2,\ldots,m\}$. The dense deployment of RRHs and mmWave communication links are the requirements of 5G. There are total n users in the network that moves freely with a certain probability. The set of users is N where $N = \{1,2,\ldots,n\}$. All the RRHs are connected to the BBU pool by the fronthaul link. The BBU pool controls the user-RRH association with the information received from users each time-stamp.

We consider a time period $\mathscr{T}$, uniformly divided into time slots t, which can be represented as $\mathscr{T} = 1,2,\ldots T$. Each User position changes at each time slot. The location co-ordinates of a user can be represented by (x_i, y_i) for for $i \in N$. The location of an RRH can be denoted by (x_j, y_j) for $j \in M$.

Table 1: Symbols and notations

m	Number of RRH
n	Number of user
M	Set of RRH
N	Set of user
(x_i, y_i)	Location of user i
(x_j, y_j)	Location of RRH j
R	RRH coverage range

$\sigma_{i,j}$	Association indicator of user i and RRH j
$D_{i,j}$	Distance between user i and RRH j
$PL(D)$	path loss
α, β, χ, o^2	path loss co-efficients
δ_i^j	SNR received by user i from RRH j
P_j	Transmit power of RRH j
Ω	Antenna gain
P_n	Noise power
θ	Angle of departure
$\mathscr{U}$	User capacity of RRH
τ_i^j	Data rate of user i connected to RRH j
δ_{th}	SNR threshold
t	Per unit time
$\mu(z)$	Fuzzy membership function
Δ_T	TTT value
F_q	Fuzzy rule
q	Number of rules
k	Index of candidate RRH
A_k	Set of candiate RRH
$\mathscr{T}$	Total time
$\mathscr{S}$	State space
$\mathscr{A}$	Action space
$\Theta_{i,j}$	Angle between user i and RRH j
$\Upsilon_{i,j}$	Direction between user i and RRH j
$x_{i,j}$	Association features
$s^t_{j,x_{i,j}}$ / s_t	State at time t

a_t	Action at time t
$X_{i,j}$	Association feature set
r_t	Reward at time t
e	Explored actions
p, q	Lagrange data points
ρ	Radius of expected region
$O_{e,h}$	Overlapping region between circles C_e and C_h
$P_{i,j}$	Proximity between user i and RRH j
$\Lambda_{i,j}$	Directional displacement between user i and RRH j
$r^k_{t,v}$	Virtual reward at time t

A. Assumptions

Several assumptions are made for the CRAN and users. We carefully considered the standard assumptions made in related works while making the assumptions. The assumptions are given below:

1. Assumptions for RRHs

The transmission range of all the mmWave small RRHs are assumed to be the same and the coverage area can be depicted by a circle with a radius R. The mmWave RRHs are equipped with a directional antenna which is necessary to provide beamforming for the mmWave system. However, the number of users an RRH can support at a certain time is limited according to the capacity of that RRH.

2. Assumptions for the BBU controller

The BBU controller exists in the BBU pool which has all information about the network. The network information is periodically updated based on the reporting of the users through the associated RRHs. The location coordinates and coverage region of all the RRHs are also known to the controller. The BBU controller runs the algorithms for carrying out handover and association decision, which is then sent to the RRHs.

3. Assumptions for user connection

Each user is assumed to be equipped with a single antenna device. It means one user can be associated to only one RRH in the network at a particular time t. The users move in the network using a modified random walk mobility model. The user is assumed to be equipped with some location service (e.g. GPS) and when a certain condition is met, the user reports its information to the serving RRH.

B. Initial user association

The association indicator between user and RRH can be denoted by $\sigma_{i,j}$ which represents whether user i is associated to RRH j or not

$$\sigma_{i,j} = \begin{cases} 1 & \text{if user } i \text{ is associated with RRH } j; \forall\ j \in M \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

Initially, the users will be associated to an RRH based on proximity. To be specific, a user will be associated to the RRH which is in the closest distance to the user. The distance between user i and RRH j can be denoted by $D_{i,j}$, which

can be calculated using Euclidean distance formula as indicated below:

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{2}$$

When the user arrives in the network, it may get signal from multiple RRHs. So, initially it is associated to the RRH which has the least distance with the user.

C. Propagation model

We assume that the channel of a mmWave RRH is based on the 3GPP standard LOS model [17]. LOS model determines that a line-of-sight mmWave link exists between the user and RRH. We have not considered the NLOS connection in the dense mmWave network with overlapping RRH based on the explanation given in [17]. According to [16], [18], the path loss model can be written as:

$$PL(D) = \alpha + 10\beta \log_{10}(D) + \chi,\ \chi \sim \mathcal{N}(0, o^2), \tag{3}$$

where D is the distance between user and RRH measured in meters, α is the floating intercept in dB, β is the linear slope over the measured distance, and o^2 is the log-normal shadowing variance. However, in mmWave band, inter-user interference can be ignored for a specific user i [16], [17]. So, we can model the signal-to-noise ratio (SNR) of the signal received by user i from RRH j as:

$$\delta_{i,j} = \frac{P_j \times \Omega \times PL(D)^{-1}}{P_n}, \tag{4}$$

where P_j is the transmit power of RRH j, P_n is the noise power, and Ω is the antenna gain. According our assumptions, RRHs are equipped with directional

antennas and users are equipped with omnidirectional antennas. So, the antenna gain Ω is a function of the angle of departure θ from the RRH to the user [17], that can be given as:

$$\Omega(\theta) = \begin{cases} \Omega_{max} & \text{if} |\theta| \leq \theta_b \\ \Omega_{min} & \text{otherwise,} \end{cases} \tag{5}$$

where Ω_{max} is the antenna gain from main lobe, Ω_{min} is the antenna gain from the side lobe, and θ_b is the width of the antenna main lobe. Beam tracking is assumed to be perfectly used for maintaining the mmWave connection between user and RRH [16]. Thus, the user can achieve a high antenna gain staying always in the main lobe.

The number of users an RRH can serve in a single time period t is limited to the capacity of that RRH. We assume that RRH j can serve $\mathscr{U}$ users simultaneously. All the users associated to RRH j is allocated the bandwidth resources evenly. So, according to Shannon capacity formula, the data rate achieved by user i connected to RRH j can be calculated by:

$$\tau_{i,j} = \frac{BW_j}{U_j} log_2(1 + \delta_{i,j}), \tag{6}$$

where BW_j is the bandwidth of RRH j and U_j is the number of users served by RRH j.

D. QoS model

In order to maintain the QoS requirement of the user with the serving RRH, we use two metrics: SNR threshold δ_{th}, and time-to-trigger (TTT) denoted by Δ_T. δ_{th} is the minimum SNR required to maintain the user-RRH connection. Δ_T is

the duration for which the user will maintain its connection while getting SNR less than or equal to the threshold value. The user waits till Δ_T becomes zero before sending the measurement report to the serving RRH. It can be said that the QoS requirement of user i is satisfied when the following condition holds

$$\exists t \in [T_c, T_{c'} - \Delta_T], s.t.\ \delta_{i,j}(t) > \delta_{th};\ \forall_{T_c, T_{c'}} \in \mathscr{T}, \tag{7}$$

where T_c and $T_{c'}$ are the time of two consecutive handovers. t is the duration for which the user gets SNR greater than the threshold value, reflecting the QoS satisfaction of user.

IV. PROPOSED HANDOVER MINIMIZATION AND USER ASSOCIATION SCHEME

In this chapter, we present the handover trigger condition and user-RRH association for the proposed framework. We first describe the handover parameter optimization method to trigger handover based on fuzzy logic. Then, we use an RL-based scheme to choose the RRH for a user during handover, which is the proposed user association scheme. According to 3GPP [7], six handover initiation events are given for cellular networks. Our main goal is to perform handover to an RRH, avoiding a short-sighted decision of choosing RRH with the highest SNR based or proximity, so that the user-RRH connection remains for a long time, thus reducing the total number of handovers.

A. Fuzzy logic-based handover parameter optimization

We implement a fuzzy logic-based method in this work for optimization of the parameter for handover triggering. The proposed method is discussed in this subsection.

1. Handover trigger condition

We consider the event A2 defined in 3GPP (Serving becomes worse than the threshold) which means when the serving RRH SNR value becomes less than the threshold SNR value. The trigger condition can be expressed as:

$$Serving\ SNR < threshold - HOM, \tag{8}$$

where HOM is a handover margin added for reducing ping-pong handover. Optimizing the value for HOM is another interesting research issue, but it is out of the scope of this paper. So we set it to be zero for simplicity.

In the traditional handover scheme, the handover event occurs when the condition of equation 8 satisfies for a predefined amount of time called TTT. Once the handover event is triggered, the user device monitors the received SNR from the serving RRH. If the received SNR does not become higher than the threshold SNR for the TTT amount of time, the user sends a measurement report to the serving RRH. The frequency of measurement report sending by the user is set by network operators [8]. The handover control parameter, TTT is necessary for minimizing early handover and late handover. High values of TTT cause too late handover, while low values of TTT lead to too early handover. So, TTT should be adjusted in a way so that the connection continues without radio link failure. In order to adjust the value of TTT, we apply fuzzy logic in this work. TTT optimization reflects the objective of our work which is maintaining the connection for as long as possible. In this part of our work, we try to maintain the connection with the current serving RRH of the user for an optimized TTT amount of time.

2. TTT optimization with Fuzzy Logic

Fuzzy logic is an inference method that maps a set of control inputs to a set of control outputs through fuzzy rules [19]. The whole process consists of three steps: fuzzification of all input values into membership functions, fuzzy reasoning based on a set of rules, and defuzzification of the output functions. The fuzzy inputs are associated with some linguistic variables. Using these linguistic

variables for each of the inputs, the rules are generated. The inference engine selects the best rule for updating the output parameter. The output determines a conclusion for each of the rules.

The main goal of using fuzzy logic here is to adjust the value of TTT when the SNR received by a user from its serving RRH becomes less than threshold SNR δ_{th}. Although, most of the handover schemes considered performing handover based on the received SNR it may lead to unnecessary and frequent handover in a small RRH-based C-RAN scenario. The RRHs are placed in a way that the coverage of some RRHs is overlapped with each other. So, a user can get SNR from multiple RRH simultaneously. It may cause ping-pong handover if the user is associated to an RRH only based on SNR. The user may get back to the previous RRH if the serving SNR becomes low at the next time period. By considering the distance of the user from its serving RRH and its velocity, we can make an approximate decision about how long the user will be inside the coverage of its serving RRH.

We fuzzify two inputs namely velocity v_i and distance $D_{i,j}$. Three linguistic variables are assigned for each of the fuzzified inputs with triangular membership functions. The triangular membership function $\mu(z)$ can be defined by a lower bound a, upper bound b, and a value m, where $a < m < b$. Each element of input x is mapped to a value between 0 and 1.

$$\mu(z) = \begin{cases} 0, & z \leq a \\ \frac{z-a}{m-a}, & a < x \leq m \\ \frac{b-z}{b-m}, & m < x < b \\ 0, & z \geq b, \end{cases} \tag{9}$$

The set of fuzzy rules contains all possible relationships among the two input values and one output value. Since each of the inputs has three linguistic variables, a total number of $(3^2) = 9$ rules are generated with all the combinations of input variables. We keep the number of linguistic variables to three as the number of linguistic variables determines the number of fuzzy rules. More fuzzy rules result in more memory requirements as well as more computation time while less number of fuzzy rules may cause inaccurate inference [20]. The output of the fuzzy process is the TTT value denoted by Δ_T. The fuzzy logic-based TTT optimization scheme is illustrated in figure 2.

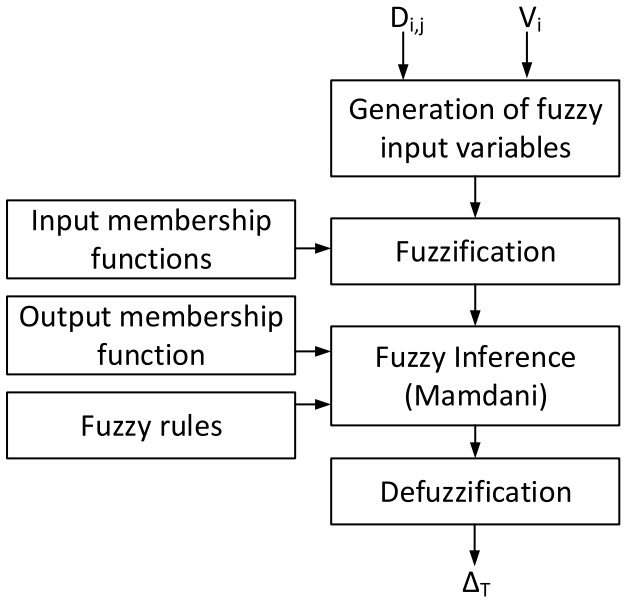

Figure 2: Fuzzy logic-based TTT optimization process

Figure 3 represents the linguistic variables of inputs with the corresponding degree of membership function as given in equation 9. As seen in the figure, velocity v_i is divided into {slow, average, and fast} and distance $D_{i,j}$ is categorized by {close, medium, and far}. We selected the core width and

boundary regions of the membership functions with a trial and error approach. It is necessary to choose the intersecting area of adjacent linguistic variables properly as more intersection causes frequent activation of multiple rules. On the other hand, less overlapping weakens the flexibility and smoothness [21]. The Mamdani-type inference method [22] is used for mapping the inputs to the output of the fuzzy system which is the value of TTT. For the TTT values, we use five sets of triangular membership functions to achieve reasonable granularity in the output: {very low, low, medium, high, and very high}. The fuzzy logic-based TTT optimization procedure is given in Algorithm 1. Initially, Δ_T is set to zero and user mobility begins. When a user meets the handover trigger condition, which means that the received SNR $\delta_{i,j}$ of user i from RRH j becomes less than or equal to the predefined threshold SNR value δ_{th}, the fuzzy rule process is called. The TTT value is updated using fuzzy rules. The TTT keeps decreasing till it becomes zero and the user keeps moving in the network with the same connection. After the end of TTT, if the received SNR condition remains, the handover event is initialized. If the received SNR becomes greater than the threshold during the time of TTT, the user is not considered for handover.

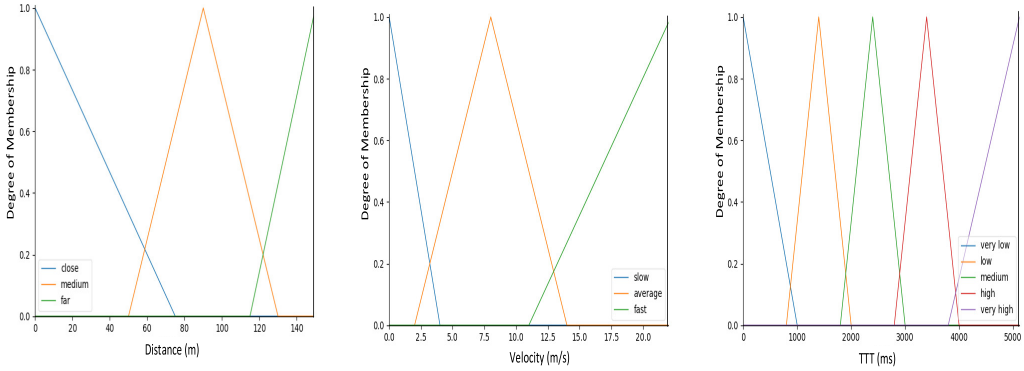

Figure 3: Membership functions of the inputs: Distance and Velocity and output: TTT

Algorithm 1 TTT optimization with fuzzy logic

Initialize SNR threshold δ_{th}, fuzzy rules, and $\Delta_T = 0$;
for time $t = 1, 2, \ldots T \in \mathscr{T}$ **do**
  for user i connected to RRH j, $\forall i$ **do**
    if $\delta_{i,j} \leq \delta_{th}$ and $\Delta_T = 0$ **then**
      Check v_i and $D_{i,j}$;
      Update Δ_T with fuzzy inference;
    else if $\Delta_T > 0$ **then**
      $\Delta_T = \Delta_T - 1$;
      if $\Delta_T = 0$ and $\delta_{i,j} \leq \delta_{th}$ **then**
        Handover event occurs;
      end if
      if $\delta_{i,j} > \delta_{th}$ **then**
        $\Delta_T = 0$;
      end if
    end if
    User moves with v_i;
  end for
end for

3. Candidate RRH selection

After the end of TTT, a suitable RRH is needed to be selected for the user. For user i sending the measurement report to the BBU pool, the BBU controller selects the candidate RRHs based on the SNR values user is receiving from the nearby RRHs. So the target RRH will be selected from only those RRHs that are selected as candidate RRHs. Let A_k be the set of available RRHs when user handover event occurs for user i at time t,

$$A_k(t) = \{k | \delta_{i,k}(t) > \delta_{th}, \forall_k \in A_k, A_k \subseteq M\}, \tag{10}$$

where k denotes the index of candidate RRHs. Our goal is to associate the user i to an RRH from set A_k for which the user-RRH connection remains longer. When the handover event occurs, the agent has to select one RRH from the candidate RRH set A_k aiming to minimize the number of handovers.

B. RL-based User Association

The RRH selection framework for user re-association is described in this subsection. When a user sends MR to the serving RRH after the end of TTT, the BBU controller chooses the appropriate RRH for the user according to the RL algorithm. The Figure 4 shows the flowchart of the fuzzy logic-based TTT optimization and RL-based RRH selection procedure.

1. Proposed RL framework

We design our RRH selection mechanism as a reinforcement learning (RL) framework. RL algorithms [23] involve an agent to learn by interacting with the environment. The agent takes an action $a_t \in \mathscr{A}$ at each decision time $t \in \mathscr{T}$ observing a state $s_t \in \mathscr{S}$, then moves to next state $s_{t+1} \in \mathscr{S}$ and receives a reward r_t as a feedback mechanism. The reward represents the objective of the problem and the aim of the agent is to maximize the overall reward. We denote the policy $\pi(s) : \mathscr{S} \rightarrow \mathscr{A}$ that maps states into actions. The goal of the agent is to learn the optimal policy π^* that maximizes the cumulative reward.

Most of the RL algorithms such as Q-learning [24] consider the reward at each iteration as a discounted reward based on the next consecutive steps. This is limited in our case as, during each handover event, future rewards do not have any impact on the current action. In our algorithm, every state is independent of

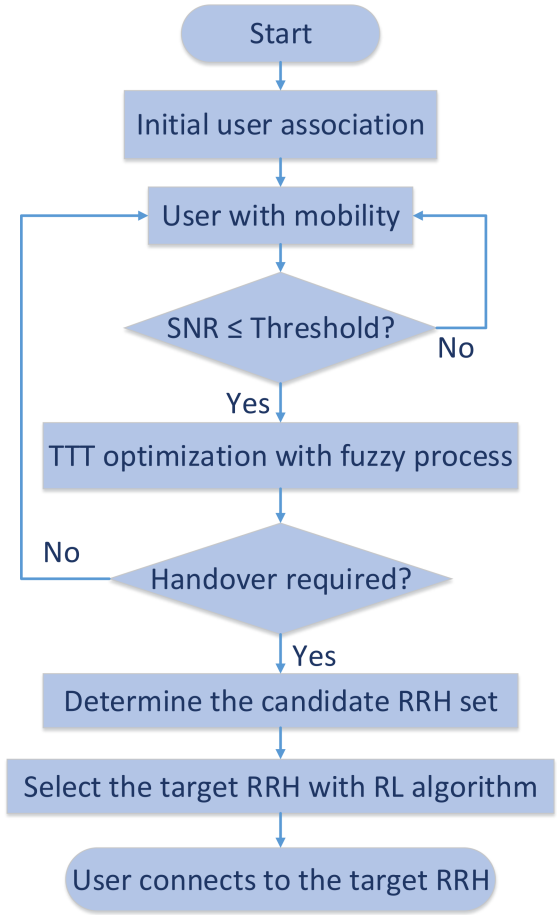

Figure 4: Flowchart of the proposed scheme

each other, and rewards received are only related to the executed action. So, the agent learns the action that often yields the best reward. This algorithm is similar to a contextual bandit framework [17]. Contextual bandits are a subset of RL algorithms that are simpler: there is only one step before the outcome is observed. The contextual bandit is an extension of the multi-armed bandit approach [25] where we factor in the context or state information. Unlike multi-armed bandit,

the state affects how a reward is associated with each action, so as states change, the model should learn to adapt its action choice. The reward is conditional to the state of the environment. To be specific, the reward is different for the same action taken at different states. The algorithm observes a context (state), takes an action from a number of available actions, and observes the outcome (reward) of that action.

In our algorithm, the candidate RRHs $k \in A_k(t)$ at each decision time t are the available actions at any particular state. When a handover event triggers following algorithm 1, the RL agent in the centralized BBU controller observes the state that includes user-RRH association information and selects an RRH by exploration or exploitation, and gets an immediate reward. Our main goal is to re-associate the user to an RRH to which the user can maintain connectivity for a longer time while satisfying the QoS requirements of the user. RL agent will learn the association of a user to an RRH based on its velocity, direction, moving angle, and distance from its associated RRH.

2. State construction

When a handover event triggers, the agent identifies the serving RRH and its association features which constructs the state of the RL agent. The state-space $\mathscr{S}$ contains four elements: serving RRH index, distance between user-RRH, angle between user-RRH, and direction of user towards RRH. At any particular state s_t at time t, the agent learns the user-RRH association information from which the handover event is triggered. So, the elements of the state can be represented by $\{j, D_{i,j}, \Theta_{i,j}, \Upsilon_{i,j}\}$. Here, j is the serving RRH index, $D_{i,j}$ is the distance between user i and RRH j, $\Theta_{i,j}$ is the angle between the user i with RRH j, and $\Upsilon_{i,j}$ is the

moving direction of i towards j.

If we combine the association features together, we can write it as $x_{i,j}$ such that $x_{i,j} = (D_{i,j}, \Theta_{i,j}, \Upsilon_{i,j})$ for any user i and RRH j. Here, x is the features of association between user i and the RRH j. $x_{i,j} \in X_{i,j}$ denotes x_{th} association feature in the total feature set. We can define the state associated with RRH $j \in M$ for any handover event requested by user i at time t denoted by $s^t_{j,x_{i,j}}$ For simplicity, we will use the notation of state at time t as s_t.

The elements of association features are continuous values. If we take all the values for these parameters, the state space may become infinite and the agent may never be able to learn the features. RL algorithms require the state space to be discrete to operate in an environment. So, it is necessary to have discrete values of the elements in the state space.

The distance $D_{i,j}$ between user and RRH is divided into five chunks such that $D_{i,j} \in \{1,2,3,4,5\}$. The smaller the values, the closer the distance between user and RRH. $D_{i,j} = 1$ means the user is in the closest distance with RRH and $D_{i,j} = 5$ is in the furthest distance with RRH. The value of angle $\Theta_{i,j}$ in the association feature of the user with the RRH is divided into eight categories, which can be given as $\Theta_{i,j} \in \{1,2,3,4,5,6,7,8\}$ where $-180^\circ \leq \Theta_{i,j} \leq 180^\circ$.

The direction $\Upsilon_{i,j}$ of i towards j is parted into two groups namely inward direction and outward direction. $\Upsilon_{i,j}$ can be calculated from the difference between the distance at time t and the distance at time $t-1$. At time t, the distance between user i and RRH j can be denoted as $D^t_{i,j}$. Similarly, at time $t-1$, the distance is $D^{t-1}_{i,j}$. If $D^t_{i,j} > D^{t-1}_{i,j}$, it indicates that the distance of the user and RRH is increased. In that case, it can be said that the user is moving in the outward direction from the RRH. Similarly, $D^t_{i,j} < D^{t-1}_{i,j}$ indicates that the user is moving in the inward direction from the RRH as the distance at the current time is less

than the distance at the previous time. $D_{i,j}^{T} = D_{i,j}^{T-1}$ means that their is no change in user movement or direction towards RRH.

3. Action

The RL agent chooses an RRH from the candidate RRH set A_k. We can denote the action at time t as $a_t \in A_k$ which is the selected RRH. The number of possible actions in the state s_t at time t is the number of available RRHs k. The exploration-exploitation policy for selecting an action is described later in Subsection 5. .

4. Reward

The reward function of the RL agent is designed to motivate the agent for taking actions that would maximize the cumulative reward. Since the objective of our work is to choose an RRH for a user which will maintain the association for the longest time, we try to design the reward to reflect our objective. In this regard, we define our reward function in state s_t for taking action a_t at time t as:

$$r_t = T_c' - T_c, \tag{11}$$

where T_c is the time when the handover occurs and the user connects to the target RRH selected by action a_t and T_c' is the next handover time. Here, t represents the iteration counter time as seconds. T_c and T_c' are assumed as the beginning and ending counter of a handover. The time unit is the same, but we use T_c and T_c' here for the convenience of representing the connection duration. So, the reward contains the duration for which the user-RRH connection remains. Maximizing the reward means the duration of connection is also maximized, thus minimizing the total number of handovers. We do not get r_t immediately after taking the

action here because we cannot calculate this until the next handover occurs.

5. Exploration-Exploitation Strategy

So, when a handover event occurs at time t, the policy is to select the RRH k^* from the candidate RRH set $A_k(t)$ satisfying

$$k^* = \underset{k}{\operatorname{argmax}} \sum_{t \in \mathscr{T}} r_t(k), \tag{12}$$

Algorithm 2 shows the overall RL-based RRH selection procedure. The algorithm is called when a handover event is triggered after the end of TTT. As described earlier, the current time is recorded as T_c. The agent observes the state s_t and checks all the available RRHs for re-association from the candidate RRH set A_k. For the ε-greedy policy, a random variable is used to determine exploration or exploitation. During exploration, the agent chooses a random RRH from the candidate RRH set and the reward is calculated. When the next handover event occurs, the time T_c' is recorded, and the total connection duration is calculated, which is the reward agent receives for taking the particular action. The exploitation procedure in this work is modified and divided into three parts based on the exploration of a state and the number of actions explored in a state.

During the exploitation phase, a virtual reward is calculated to choose the best action in two cases: the state was not explored before, and some actions of the state are explored. When the agent gets in a state which was not explored before such that $s_t \notin \mathscr{S}$ or the agent gets in a state for which only some actions $e \in A_k$ were explored, the virtual reward is calculated. It is calculated for all the available actions k in a state s_t based on a future location prediction mechanism. This mechanism is used for faster convergence of the algorithm which is called

the acceleration technique in this work.

When the first condition satisfies such that the state s_t is a new state, the agent takes the action which has the maximum virtual reward defined as $r^k_{t,v}$. In the second case, the agent calculates the virtual reward for all the available actions similarly. Then, a bias value b is calculated using the actual reward and virtual reward for the explored actions $e \in A_k$ in state s_t. After that, for all the explored actions, the new reward $r^{e'}_t$ is calculated for the unexplored actions $\forall e' \in A_{e'}$, by multiplying the bias value with the virtual reward values. Then for all the explored and unexplored actions, the action with the highest reward is selected by the agent. Here, reward means both the actual reward for explored actions r^e_t and the new calculated reward $r^{e'}_t$. Lastly, if all the available actions in a state s_t are explored before, the agent selects the action with the maximum reward.

Algorithm 2 RL-based RRH selection algorithm

Initialize ε, total simulation time $\mathscr{T}$
while Handover event occurs according to Algorithm 1 **do**
    Record the current time $T_c \in \mathscr{T}$
    Observe the state s_t
    Check all the available actions k where $k \in A_k; A_k \subseteq M$
    if Rand(0,1) $< \varepsilon$ /**Exploration**/ **then**
        Select a random action $a_t \in A_k$
        Observe the reward r_t according to Equation 11 when the next handover occurs at $T_c' \in \mathscr{T}$
    else /**Exploitation**/
        if $s_t \notin \mathscr{S}$ **then**
            Calculate the virtual reward $r_{t,v}^k, \forall_k \in A_k$
            Select action $a_t = argmax_k\ r_{t,v}^k$
        else if Some actions $e \in A_k$ are explored **then**
            Calculate the virtual reward $r_{t,v}^k, \forall k \in A_k$
            Calculate bias $b = \frac{r_t^e}{r_{t,v}^e}, \forall e \in A_k$
            Calculate the new reward $r_t^{e'} = b * r_{t,v}^{e'}, \forall e' \in A_{e'}$ where $A_{e'} = A_k \backslash \{e\}, \forall e$
            Select action $a_t = argmax_{e,e'}(r_t^e \cup r_t^{e'}), \forall e, e' \in A_k$
        else
            Select action $a_t = argmax_{a_t}(r_t)$
        end if
    end if
end while

6. Acceleration technique

In order to calculate the virtual reward, we use a prediction method with Lagrange-based extrapolation using the past trajectory of the user. Then, we utilize the approximated future location of the user to create an overlapping region with the RRHs. The future overlapping region reflects how long the user may stay under the coverage of a certain RRH. So, the agent updates the virtual

reward for the corresponding state-action pair based on the overlapping region values and other two parameters namely proximity and direction.

- Future location prediction

  Lagrange polynomial is a useful method for producing an approximation for an arbitrary function [26]. The location is calculated in a two-dimensional space with respect to time. Using the past location coordinates of a user from few consecutive time stamps, we utilize the extrapolation capability of the Lagrange method to get the future location of the next timestamp. At time t, the location of user i with respect to time can be denoted by $(X_{i,t}, Y_{i,t})$. We determine the location co-ordinates of the user for $n+1$ number of data points for degree n, where $n = 1, 2, \ldots, t-1$. We can generate the X-axis and Y-axis value with respect to time separately, and then determine the future location of a user through the extrapolated values. The future location of user i at time t' for degree n with X-axis value $X_{i,t'}$ and Y-axis value $Y_{i,t'}$ can be calculated by the following

$$X_{i,t'} = \sum_{p=1}^{n} \left[X_p \prod_{\substack{q=1 \\ q \neq p}}^{n} \frac{t' - t_q}{t_p - t_q} \right], \tag{13}$$

  and

$$Y_{i,t'} = \sum_{p=1}^{n} \left[Y_p \prod_{\substack{q=1 \\ q \neq p}}^{n} \frac{t' - t_q}{t_p - t_q} \right], \tag{14}$$

  where p and q are the values of the data points in consecutive time stamps. t' is the time for which the future location of the user will be approximated. The

future location of user i for the next time stamp t' can be denoted by $(X_{i,t'}, Y_{i,t'})$.

- Overlapping region creation

  From the predicted location $(X_{i,t'}, Y_{i,t'})$, we have created an expected region based on the velocity of the user. The expected region is a circle C_e which includes all the possible locations the user can be in the few consecutive future time stamps. The expected region circle C_e is created for user i centering the future location $(X_{i,t'}, Y_{i,t'})$ with radius ρ which can be calculated by the predicted displacement of the user given as:

$$\rho = \sqrt{(X_{i,t'} - X_{i,t})^2 + (Y_{i,t'} - Y_{i,t})^2}, \tag{15}$$

  where t' is the future time stamp for when the location is approximated, and t is the current time stamp. We calculate the overlapping region between the expected region circle and the RRH coverage range circle. The overlapping region between a user and an RRH determines how long the user may stay under the coverage of that RRH.

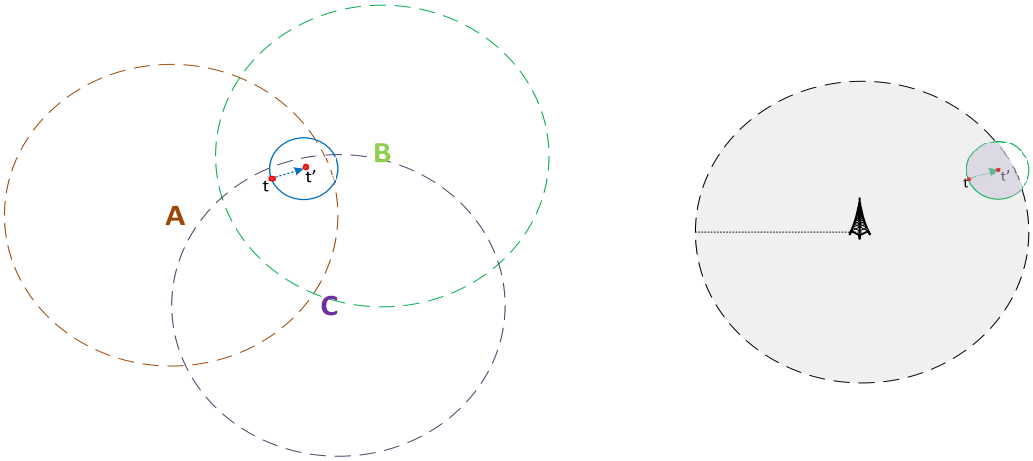

Figure 5: (a) Expected region creation with the predicted location (b) Overlapping region between the expected region circle and RRH coverage circle

To get the overlapping region between two circles C_e and C_h denoted by $O_{e,h}$ such that $O_{e,h} = Area(C_e) \cap Area(C_h)$, we need the radius of both circles and the distance between the centre of two circles denoted by d_c.

We can calculate the overlapping region $O_{e,h}$ by the following equation

$$O_{e,h} = \frac{\rho^2(\Phi_\rho - \sin\Phi_\rho) + R^2(\Phi_R - \sin\Phi_R)}{2}, \tag{16}$$

where $0 \leq O_{e,h} \leq 1$, and the values of Φ_ρ and Φ_R can be calculated by the following formulas

$$\Phi_\rho = 2\cos^{-1}\left(\frac{d_c^2 + \rho^2 - R^2}{2\rho d_c}\right), \tag{17}$$

and

$$\Phi_R = 2\cos^{-1}\left(\frac{d_c^2 + R^2 - \rho^2}{2R d_c}\right), \tag{18}$$

where ρ is the radius of the expected region circle C_e and R is the radius of the

RRH coverage range circle C_h. As $C_e < C_h$, it may occur that C_e is completely inside C_h. In such case, the overlapping area will be equal to $Area(C_e)$ which can be calculated by

$$Area(C_e) = \pi\rho^2, \tag{19}$$

where ρ is the radius of circle C_e.

- Virtual reward calculation

Maximizing the overlapping region $O_{e,h}$ between the expected region of user i and coverage range of RRH j at time t is equivalent to maximizing the duration of the user-RRH association. This overlapping region is used for calculating the virtual reward $r^k_{t,v}$ for all the available actions k at time t when a certain condition of exploitation occurs. Along with that, we include the proximity of user i and RRH j, and the directional displacement of the user in the virtual reward function. The proximity can be calculated by

$$P_{i,j} = \frac{D_{i,j}}{R}, \tag{20}$$

where $D_{i,j}$ is the distance between user i and RRH j, and R is the RRH coverage range. This proximity determines how close user i is with the RRH j, which means higher proximity indicates the user is closer to an RRH. The directional displacement here is related to the direction $\Upsilon_{i,j}$ calculated in the state space in Subsection 2. . The directional displacement of user i towards RRH j can be calculated by

$$\Lambda_{i,j} = \frac{D^{t-1}_{i,j} - D^{t}_{i,j}}{v_i}, \tag{21}$$

where v_i is the velocity of user i. The positive value of $\Lambda_{i,j}$ indicates that the

user i is moving towards RRH j, and the negative value indicates the user is moving in the outward direction. Maximizing the value of proximity and moving direction along with the overlapping region ensures that the user has more possibility to stay under that RRH for a longer time.

So, at every decision time t, the virtual reward for each candidate RRH can be calculated by

$$r_{t,v}^{k} = \begin{cases} -1 & \text{if } O_{e,h}^{t} == 0,\ \forall k \in A_k \\ O_{e,h}^{t} + P_{i,j} + \Lambda_{i,j} & \text{otherwise.} \end{cases} \tag{22}$$

This virtual reward is mapped with the actual reward to calculate a bias value b. The bias is used to calculate the new reward for certain exploitation phases as given in Algorithm 2.

V. PERFORMANCE EVALUATION

In this section, we evaluate the proposed fuzzy logic-based handover triggering and RL-based user association scheme with acceleration technique (FLRL_ac) in various scenarios. Theoretically, the cumulative reward is the common metric to evaluate an RL algorithm. However, in our case, the optimal RRH selection decision is difficult to compute due to the large scale of the problem. Therefore, in order to evaluate the performance of our proposed scheme, we compare it with the traditional SNR-based handover (SBH) scheme. In addition to that, we also verify the performance of FL-based TTT-optimization and the acceleration technique of RL separately. In this regard, we implemented two schemes namely RL_ac without FL and FLRL without ac.

The SBH selects the RRH for user association based on the highest SNR. We kept the handover trigger condition the same as that of our work. RL_ac is the RL-based RRH selection with acceleration technique without the fuzzy logic. The TTT is not optimized in RL_ac. In FLRL without ac, the acceleration technique is not used for RL-based RRH selection. The fuzzy logic-based TTT optimization is used here same as the FLRL_ac.

A. Simulation Environment

We consider a C-RAN environment that covers a 1000(m) X 1000 (m) square region and consists of a certain number of small RRHs randomly deployed. The coverage range of all the RRHs is same and overlaps with other neighboring RRHs, each represented by a circular area. The number of RRHs is set to 50 by default. The transmit power of RRH is set to 30 dBm and the noise power is -77 dBm. The parameters for path-loss calculation in Equation 3 are similar to [18] corresponding to a carrier frequency of 28 GHz and LOS communication.

The bandwidth allocated to RRH is set as 500 MHz. The users are randomly distributed in the simulation area move in the network with a modified random walk model. The default number of users and user velocity are 200 and 6(m/s). The simulation parameters used in this work are summarized in Table 2.

Table 2: Simulation parameters

Parameters	**Values**
Size of network area	(1000 X 1000) m
RRH transmit power	30 dBm
Noise power	- 77 dBm
Bandwidth	500 MHz
Parameters for path loss	$\alpha = 61.4,\ \beta = 2$
RRH coverage range	150 m
Number of RRH	50 (default)
Number of user	200 (default)
User capacity of RRH	10
Number of iterations	10000
Epsilon (ε)	[1, 0.1, 0.99]

The layout of the network is depicted in Figure 6 with 40 RRHs and 200 users. The black lines indicate the coverage range of each RRH and red circles represent the users in the network. The straight blue lines indicate the walking path of the user. We have assumed that the user can move only through the straight lines with a modified random walk.

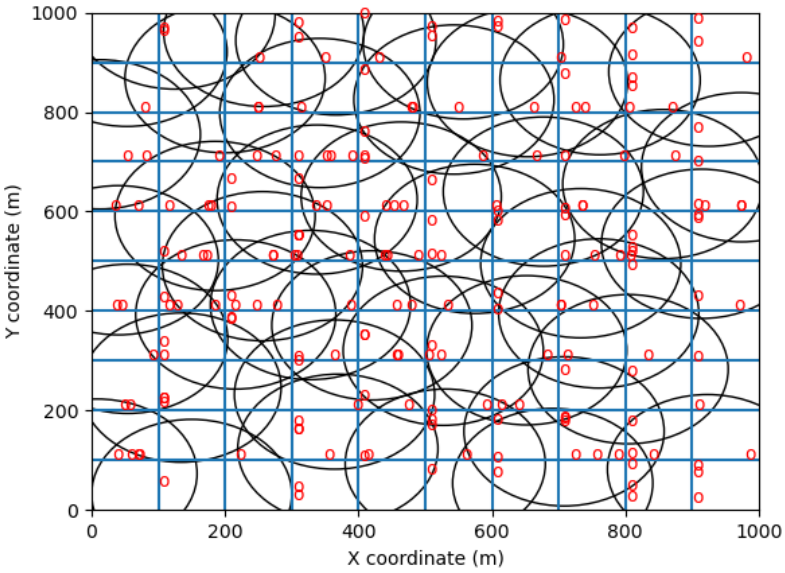

Figure 6: The layout of the network with 40 RRHs and 200 users

B. Numerical Results and Discussions

The performance of our proposed scheme is evaluated by varying different parameters. This subsection discusses the evaluation results with different parameters in terms of the number of handovers per user and average reward by comparing with different schemes. Average reward represents the average connection remaining time for the user-RRH association. As our objective is to maintain the connection duration longer and reduce the number of handovers while maintaining QoS, these two metrics can reflect the performance of our proposed scheme over the compared schemes.

1. Convergence evaluation

We start by analyzing the convergence of FLRL_ac with only FLRL without acceleration technique. The main reason is to show the benefit of using the proposed acceleration technique. To do so, we show the total number of handovers and average reward with the increasing number of episodes. We kept the default network parameters and run the two schemes for 100,000 iterations for the simulation.

Figure 7 shows the converge of the RL algorithms with the total number of handovers per 10,000 episodes. It can be observed that both algorithms reach convergence by 50,000 episodes. Although both of them converged, the performance of the FLRL_ac in terms of the total number of handovers per user is better than that of FLRL without ac.

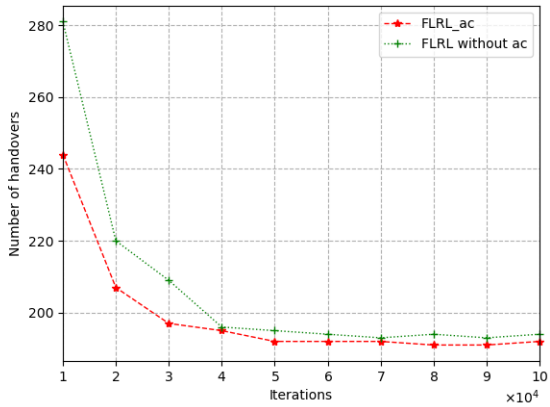

Figure 7: Performances on convergence of the RL algorithm with number of handovers

The performance comparison on the convergence of the RL algorithm in terms of the average reward is shown in Figure 7. The average reward is the average duration of the user-RRH association. It can be seen from the figure that

FLRL_ac outperformed FLRL without ac in terms of average reward.

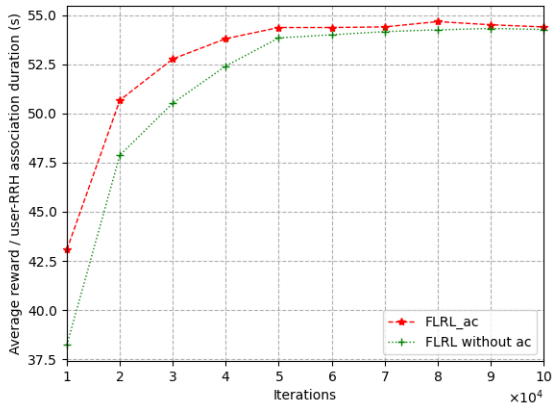

Figure 8: Performances on convergence of the RL algorithm with average reward

2. Varying density of RRHs

We choose eight values for the number of RRH: 30, 40, 50, 60, 70, 80, 90, and 100 and run 10000 iterations (time unit) for each instance. We examine the number of handovers per user while keeping the number of users and user velocity as default. The results are shown in Figure 9.

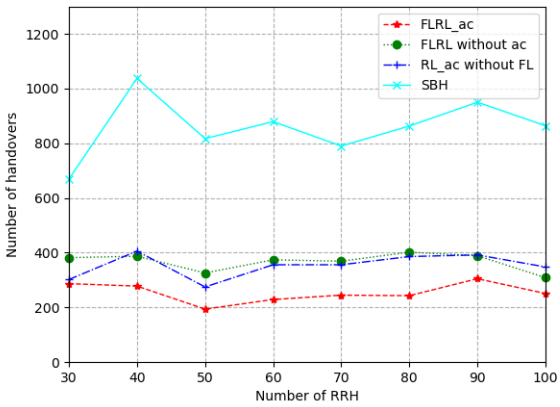

Figure 9: Performance on the number of handovers with varying number of RRHs

From the figure, it can be seen that the number of handovers for FLRL_ac is significantly smaller than that of the FLRL without ac, RL_ac without FL, and SBH. The benefit of using both the Fuzzy logic-based TTT optimization and acceleration technique can be realized from this result. The number of handovers is the lowest when the number of RRH is 50 and it increases slightly when the density becomes higher than 50 to 90 in our C-RAN environment. The density of RRH has an impact on the number of handovers, because in the same region when a certain number of RRHs are deployed, the RL agent has more options to choose the best RRH to reduce the overall number of handovers.

We keep the RRH density and other parameters the same and compare the average user-RRH association duration of our proposed scheme. The result is shown in Figure 10.

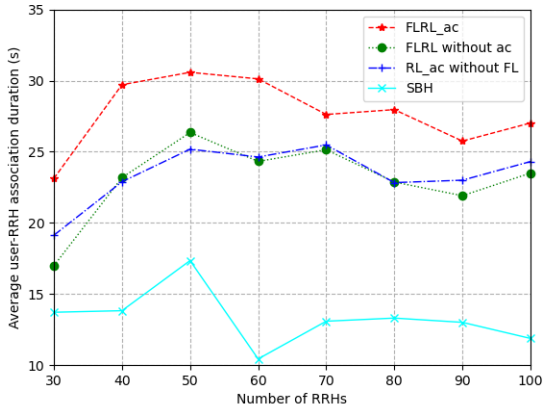

Figure 10: Performance on the average user-RRH association duration with varying number of RRHs

It can be found that FLRL_ac outperforms all the other compared schemes in terms of average user-RRH association duration. When the number of RRHs increases from 30 to 50, the duration increases and it again decreases with increasing the number of RRHs. This is because, when there are 30 RRHs, the user moves under the coverage of less number of RRHs and the agent may select an RRH for that user may not stay longer. Again, the duration starts decreasing for more than 50 RRHs due to the exploration period of the agent, when the agent chooses different RRHs and learns the reward. The candidate RRH set becomes larger and the agent takes a longer time to converge to the best action.

3. Varying number of users

We vary the number of users in our C-RAN environment to verify the performance of our algorithm in terms of the number of handovers and average user-RRH association duration. We vary the number of users to 100, 150, 200, 250, 300, 350, and 400 in the default network setting.

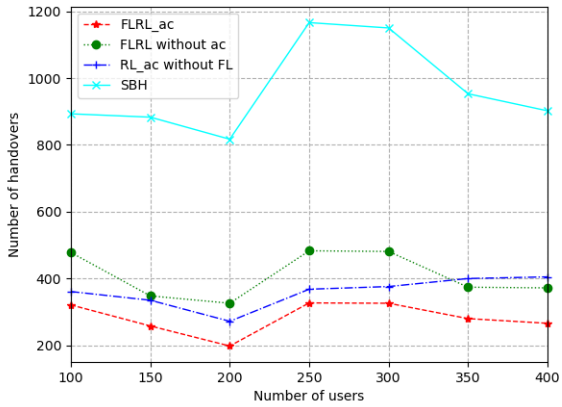

Figure 11: Performance on the number of handovers with varying number of users

The performance on the number of handovers is given in Figure 11. It can be observed that FLRL_ac outperformed other algorithms in terms of the number of handovers per user with varying number of users. Although the number of handovers for RL_ac without FL was lower than that of FLRL without ac in the beginning, when the number of users increases to 350, the number of handovers slightly increases.

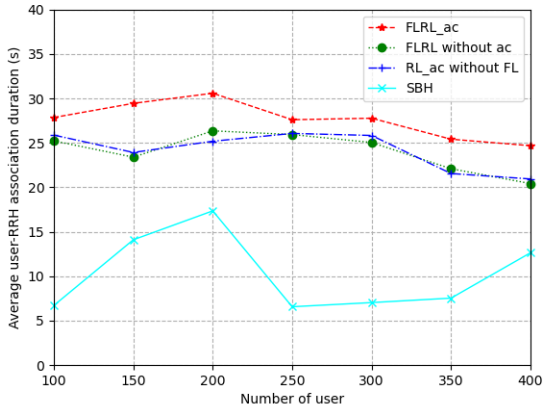

Figure 12: Performance on the average user-RRH association duration with varying number of users

The performance of the average user-RRH association is displayed in Figure 12. FLRL_ac outperforms all the other compared schemes and the average duration is the highest in the default settings when the number of user is 200. The performance decreases slightly with increasing the number of users to more than 200 in the network. The performance of FLRL and RL_ac is almost similar with varying number of users.

4. Varying user velocity

The velocity of the user has a significant impact on the performance of the proposed method. The handover control parameter directly depends on the velocity of the user. So, we investigate the performance of the proposed scheme with different velocities of users.

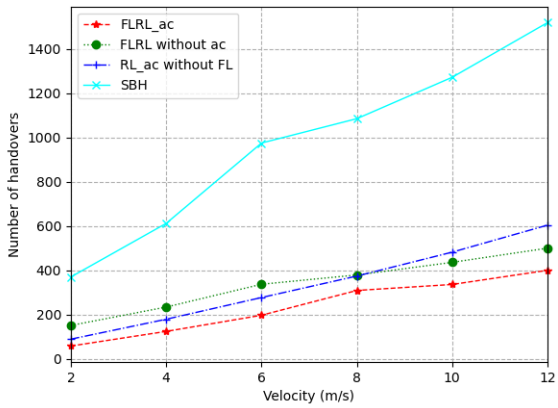

Figure 13: Performance on the number of handovers with varying user velocity

The result of the number of handovers per user is shown in Figure 13. FLRL_ac performs better than the other compared schemes in terms of the number of handovers per user. The number of handovers for RL_ac without FL was less than that of FLRL without ac at the beginning. When the velocity increases, the number of handovers increases for RL_ac because it does not directly consider the user velocity. Since handover triggering is performed with FL based on user-RRH distance and user velocity, the TTT is optimized for both FLRL_ac and RL_ac with the increasing velocity.

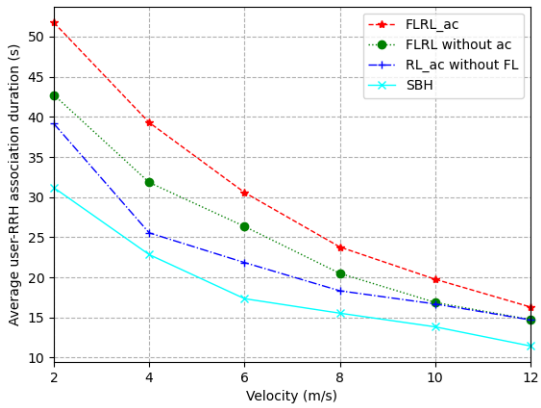

Figure 14: Performance on the average user-RRH association duration with varying user velocity

Figure 14 shows the average user-RRH association duration with varying user velocity. It can be observed that the association duration decreases with increasing user velocity. As the velocity increases, the user moves away from the coverage region of an RRH very quickly. Also, the received SNR becomes low very fast. So, the handover condition triggers and target RRH is selected by the RL agent when all the conditions are satisfied. It can be observed that FLRL_ac outperforms all the other schemes because of TTT optimization and acceleration techniques.

VI. CONCLUSION

In this paper, we have studied the user re-association problem in small-RRH based C-RAN for reducing the number of handovers while maintaining the QoS requirements of users. In order to decrease frequent handovers, we have optimized the handover trigger condition as well as RRH selection for the users. At first, we have implemented a fuzzy logic-based solution for adjusting the amount of time to maintain a connection with the serving RRH after a certain threshold is met. When the handover event is triggered, an RL-based algorithm is proposed for selecting an RRH such that the connection stays longer. For faster convergence of the RL algorithm, an acceleration technique is proposed based on the prediction of users' future location. We have solved the exploration-exploitation trade-off in RL by providing a virtual reward in each RRH selection period. A mapping between the virtual reward and actual reward is performed to take the RRH selection decision in uncertainty. It has been shown that incorporating the virtual reward leads to faster convergence of the RL algorithm. In the future, we plan to extend this work by incorporating the load balancing between RRHs along with reducing the number of handovers while maintaining user demand. Moreover, we will extend the network scenario for user re-association in the heterogeneous C-RAN by including macro RRH.

PUBLICATIONS

A. Journals

1. R. T. Rodoshi, T. Kim, and W. Choi, "Resource management in cloud radio access network: Conventional and new approaches," *Sensors*, vol. 20, no. 9, p. 2708, 2020.

1. R. T. Rodoshi and W. Choi, "A survey on applications of deep learning in cloud radio access network," *IEEE Access*, vol. 9, pp. 61 972–61 997, 2021. DOI: 10.1109/ACCESS.2021.3074180.

B. Conferences

1. R. T. Rodoshi, S. Shin, and W. Choi, "A survey on deep learning for cloud radio access networks," in *Proc. of 9th International Conference on Smart Media and Applications (SMA 2020)*, 2020.

1. R. T. Rodoshi, T. Kim, and W. Choi, "Deep reinforcement learning based dynamic resource allocation in cloud radio access networks," in *Proc. Of 2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 2020, pp. 618–623. DOI: 10.1109/ICTC49870.2020.9289530.

1. R. T. Rodoshi and W. Choi, "Accuracy analysis of user location prediction using lagrange polynomial," in *Proc. of Symposium of the Korean Institute of communications and Information Sciences*, 2020, pp. 47–48.

REFERENCES

[1] A. Gupta and R. K. Jha, "A survey of 5g network: Architecture and emerging technologies," *IEEE access*, vol. 3, pp. 1206–1232, 2015.

[2] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmwave) for 5g: Opportunities and challenges," *Wireless networks*, vol. 21, no. 8, pp. 2657–2676, 2015.

[3] M. H. Alsharif, R. Nordin, M. M. Shakir, and A. M. Ramly, "Small cells integration with the macro-cell under lte cellular networks and potential extension for 5g," *Journal of Electrical Engineering & Technology*, vol. 14, no. 6, pp. 2455–2465, 2019.

[4] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks—a technology overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2014.

[5] R. T. Rodoshi, T. Kim, and W. Choi, "Resource management in cloud radio access network: Conventional and new approaches," *Sensors*, vol. 20, no. 9, p. 2708, 2020.

[6] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, 2014.

[7] E. U. T. R. Access, "Radio resource control (rrc)," *Protocol specification (Release 10)*, vol. 290, 2013.

[8] D. Castro-Hernandez and R. Paranjape, "Optimization of handover parameters for lte/lte-a in-building systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 5260–5273, 2017.

[9] K.-L. Tsai, H.-Y. Liu, and Y.-W. Liu, "Using fuzzy logic to reduce ping-pong handover effects in lte networks," *Soft Computing*, vol. 20, no. 5, pp. 1683–1694, 2016.

[10] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alriah, and A. B. Abas, "Advanced handover self-optimization approach for 4g/5g hetnets using weighted fuzzy logic control," in *2019 15th International Conference on Telecommunications (ConTEL)*, IEEE, 2019, pp. 1–6.

[11] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, and A. Alquhali, "Velocity-aware handover self-optimization management for next generation networks," *Applied Sciences*, vol. 10, no. 4, p. 1354, 2020.

[12] K. C. Silva, Z. Becvar, E. H. Cardoso, and C. R. Francês, "Self-tuning handover algorithm based on fuzzy logic in mobile networks with dense small cells," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, IEEE, 2018, pp. 1–6.

[13] M. S. Mollel, A. I. Abubakar, M. Ozturk, S. Kaijage, M. Kisangiri, A. Zoha, M. A. Imran, and Q. H. Abbasi, "Intelligent handover decision scheme using double deep reinforcement learning," *Physical Communication*, vol. 42, p. 101 133, 2020.

[14] T. Kim, C. Chun, and W. Choi, "Optimal user association strategy for large-scale iot sensor networks with mobility on cloud rans," *Sensors*, vol. 19, no. 20, p. 4415, 2019.

[15] T. Bilen, B. Canberk, and K. R. Chowdhury, "Handover management in software-defined ultra-dense 5g networks," *IEEE Network*, vol. 31, no. 4, pp. 49–55, 2017.

[16] Y. Sun, G. Feng, S. Qin, Y.-C. Liang, and T.-S. P. Yum, "The smart handoff policy for millimeter wave heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1456–1468, 2017.

[17] L. Sun, J. Hou, and T. Shu, "Spatial and temporal contextual multi-armed bandit handovers in ultra-dense mmwave cellular networks," *IEEE Transactions on Mobile Computing*, 2020.

[18] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[19] T. J. Ross *et al.*, *Fuzzy logic with engineering applications*. Wiley Online Library, 2004, vol. 2.

[20] G. Głowaty, "Enhancements of fuzzy q-learning algorithm," *Computer Science*, vol. 7, no. 4, p. 77, 2005.

[21] F. Pervez, M. Jaber, J. Qadir, S. Younis, and M. A. Imran, "Memory-based user-centric backhaul-aware user cell association scheme," *IEEE Access*, vol. 6, pp. 39 595–39 605, 2018.

[22] I. Iancu, "A mamdani type fuzzy logic controller," *Fuzzy logic-controls, concepts, theories and applications*, pp. 325–350, 2012.

[23] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[24] C. J. Watkins and P. Dayan, “Q-learning,” *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[25] S. Boldrini, L. De Nardis, G. Caso, M. T. Le, J. Fiorina, and M.-G. Di Benedetto, “Mumab: A multi-armed bandit model for wireless network selection,” *Algorithms*, vol. 11, no. 2, p. 13, 2018.

[26] D. B. Clifton, *Technique for approximating functions based on lagrange polynomials*, US Patent 6,976,043, 2005.

ACKNOWLEDGEMENTS

I would like to express my thankfulness to all the people that have helped me in the time of completing my Master's degree and research.

First of all, I would like to express my immense respect and gratitude to my advisor, Prof. Wooyeol Choi for providing me with the opportunity to pursue my Master's degree at Chosun University. His constant encouragement, support, and helpful suggestions have guided and motivated me during tough times in my study and research. His supervision and persistent guidelines have helped me towards performing good research works. I would always be thankful to him for the valuable lessons of professionalism, time management, and discipline.

Secondly, I wish to express my warm and sincere thanks to the thesis committee members, Prof. Seok Joo Shin and Prof. Moon Soo Kang for their constructive comments and helpful suggestions. All their insights regarding my thesis have helped me in refining and expanding it from various perspectives.

Thirdly, I am grateful for the opportunity to be a part of such a diverse group of students, faculty, and staff in the Department of Computer Engineering, Chosun University. Also, I would like to express my sincere acknowledgment to Smart Networking Lab for providing me with such a great opportunity and an atmosphere to learn academically and otherwise. I am thankful to my lab-mates for their moral as well as academic support. Furthermore, I would like to show my gratitude to all my seniors and friends from Bangladesh at Chosun University for their kindness and cooperation that made my life in South Korea easier and enjoyable.

Last but not the least, I would like to thank my parents, family members, and friends for their unconditional and continuous support through my difficult times.

Without their inspiration and guidance, it would have been impossible for me to achieve anything.